基于标签与关系网络的用户聚类推荐研究*

熊回香　蒋武轩
(华中师范大学信息管理学院　武汉 430079)

摘要:【目的】利用用户标签及关系网络, 为用户推荐潜在的相似用户。【方法】通过探究社会化标注系统中标签、关系网络所表征的用户长短期兴趣特征, 综合用户标签及关注关系, 利用多维尺度法构建用户聚类模型, 根据用户聚类结果进行相似用户推荐, 并以"微博"为例对模型进行实证。【结果】实验结果表明, 基于标签和关系网络的用户聚类模型能够有效地结合用户长短期兴趣特征, 挖掘潜在相似用户, 聚类及推荐效果较好。【局限】样本数据集具有局限性, 不能完全涵盖用户兴趣领域, 仅从一个领域验证了模型的准确性与有效性。【结论】通过对用户标签及关系网络挖掘用户长短期兴趣, 构建的基于用户静态标签与动态关系网络的用户推荐模型, 对个性化用户推荐效果有较好的提升。

关键词: 社会化标注　标签　关系网络　用户聚类　多维尺度分析
分类号: TP181

1　引　言

社会化标注又叫协同标注、大众分类等, 是指由网络用户自发地定义一组标签描述某类信息, 并选用高频标签作为该类信息类名的一种网络信息分类方法[1]。随着信息技术的快速发展, 国内外出现了大批允许用户自行创建标签的社会化标注系统, 如YouTube、微博等[2]。但由于用户创建标签时的随意性产生的问题, 如标签歧义、模糊、冗余等, 降低了内容标引和检索的有效性[3]。因此, 如何提高社会化标注系统信息推荐的准确性, 解决用户获取信息困难成为研究和关注的重点[4]。目前, 主流的解决方式是利用聚类算法根据用户信息对用户进行相似度计算, 实现用户聚类, 再根据用户聚类结果在同簇用户之间进行信息推荐[5], 即用户聚类结果是社会化标注系统信息推荐的依据。

(1) 社会化标注系统的推荐研究主要集中于根据用户"标签–资源"关系对相似用户进行发现, 极少将用户关系网络考虑其中, 如易明等[6]和王向前等[7]通过 VSM 将标签表示成 Web 资源向量的形式, 进而计算标签间的相似度, 利用 DBSCAN 实现标签的聚类; Gemmell 等[8-9]同样使用 VSM 构建标签与 Web 资源间的标注关系, 利用层次聚类获取标签的聚类结果并将其应用到标签的个性化推荐中。

(2) 在社会化标注领域中多维尺度分析(MDS)方法在国内主要应用于通过科学图谱以发现词间关系, 还未将其应用到相似度计算中, 如卢小宾等[10]借助 MDS 和聚类可视化分析方法构建科学图谱, 对社会化标签研究领域中的热点词汇进行识别, 揭示这些热点关键词之间的亲疏远近关系; 柴彦[11]通过 SPSS 软件的聚类分析以及多维尺度分析, 研究关键词之间的内在联系, 探究知识管理领域中的研究热点。国外已经将 MDS 应用于相似度计算领域, 如 Masnick 等[12]利用 MDS 创建职业相似性的空间表示, 用于衡量学生对职业的态度, 以鼓励学生从事科研领域的相关工作。

因此, 本文提出将标签和关系网络两者结合以挖掘潜在相似用户, 并运用 MDS 方法对表征用户长期静态兴趣的标签和用户短期动态兴趣的关系网络进行矩阵降维以计算相似度, 通过聚类寻找出兴趣和关注

通讯作者: 熊回香, ORCID: 0000-0001-9956-3396, E-mail: 648917179@qq.com。
*本文系国家社会科学基金项目"大众分类中标签间语义关系挖掘研究"(项目编号: 12BTQ038)的研究成果之一。

相似度最高的用户群体，从而实现用户的个性化推荐。同时由于用户的兴趣随着时间不断变化，不同时间用户兴趣也会有所不同，但标签的变化周期较长，具有一定的稳定性，而关系网络变化周期短，具有动态性。模型通过不断更新用户的关注变化信息以修正推荐结果，有效地解决了推荐系统的数据稀疏性，但无法兼顾用户长短期兴趣及推荐准确性等问题。经过实证研究后发现将用户关注加入到用户聚类指标中，不仅大大增强了用户聚类的准确度，而且能够揭示标签的语义关联。

2 模型描述及数据预处理

本文选取国内社会标注网站的微博数据作为实证研究的对象，微博是一种通过关注机制分享简短实时信息的广播式的社交网络平台[13]。微博用户关系的形成是在现实有联系的基础上加以个人兴趣为导向的自组织拓扑体系。对用户进行个性化推荐的核心和关键就是挖掘用户个人兴趣和偏好，为了能够准确地挖掘微博中存在的不同兴趣用户群体，可以通过构建完善的用户兴趣发现模型，在计算出用户间兴趣相似度的基础上进行聚类，在聚类簇群的基础上对用户进行精准的个性化推荐。

2.1 用户聚类模型总体框架

微博是以用户兴趣和关注关系为导向的用户关系结构和组织方式，本文整合这两种因素，在传统基于静态标签构建用户兴趣模型的基础上将用户动态关注关系这一指标引入其中并构建用户推荐模型，模型包含两个子模型：用户标签模型与用户关注模型。从而计算出稳定的相似用户群体进行聚类，提高了用户个性化推荐的效率和准确率，如图 1 所示。

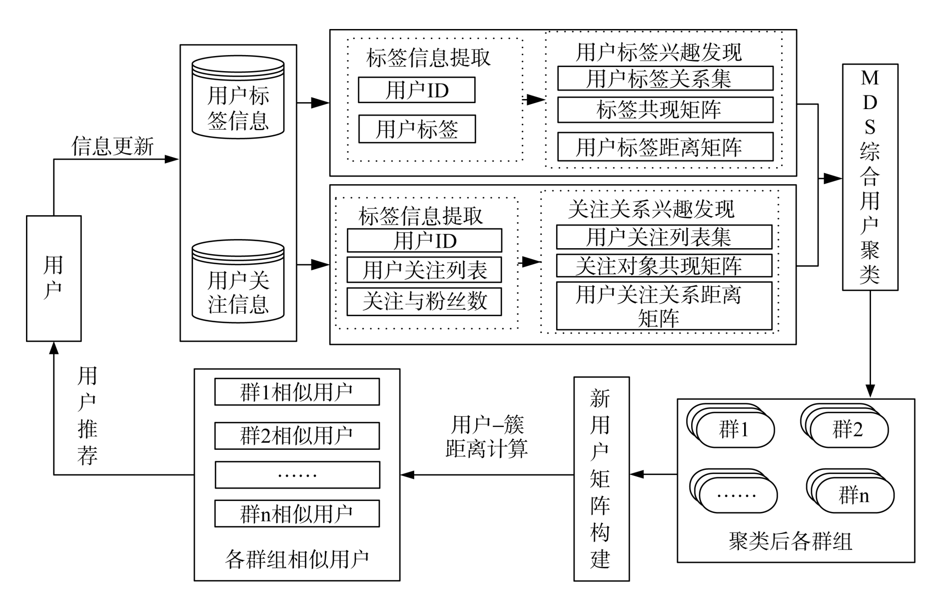


图 1 用户聚类模型框架

模型自动从微博中收集用户相关信息存入数据库中，对用户标签信息、关注信息进行信息提取，并依据模型进行数据预处理，分别生成用户标签共现矩阵及用户关注对象共现矩阵，根据共现矩阵分别计算基于标签和关注关系的用户间距离矩阵，再通过 MDS 降维将用户标签及关注关系的复杂距离整合形成二维数据，进而对用户进行聚类，实现用户推荐。同时，在较短周期内不断更新用户关注信息，不断修正用户聚类结果。这样聚类得到的结果才能够更加准确地反映当前的现实状况。

2.2 实验数据

(1) 数据获取

实证数据来自新浪微博用户数据，笔者于 2016 年 11 月 5 日利用 Python 爬虫从微博选取一名用户 (http://weibo.com/u/3660593213?from=myfollow_all) 开始逐步扩散抓取用户信息，共抓取 1 075 名微博用户，其中共有 341 名用户编辑了 1 905 个标签，表 1 显示了部分用户数据。数据集中的字段分别为：用户 ID、用户昵称、微博数、关注数、粉丝数、标签、关注列表。

表 1　部分微博用户数据

用户 ID	用户昵称	微博数	关注数	粉丝数	标签	关注列表
3694919990	各国美食学起来 YOU	102 390	118	986 725	新闻趣事, …微博奇葩	1857414070, …
5590998575	不懂老兮	806	41	532 314	外貌协会, …星座运势	3725773862, …
3323442082	视觉酱	100 402	238	2 478 436	教育就业, …时尚	3193150774, …
2155768741	贵州旅游广播	3 667	248	316 615	FM972, …快乐	2760471402, …
3524931687	走走客云南旅游	271	137	60	云南旅游, …自驾旅游	3273935392, …
1990226474	昆宣发布	28 722	1 023	621 450	春城艺术, …春城人物	1266286555, …
3175953062	萌萌萌熊	55	9	759	时尚, …星座命理	1642909335, …
…	…	…	…	…	…	…

(2) 数据预处理

①删除不完整数据

由于用户数据是通过爬虫自动抓取的，因此存在一些抓取不完整的现象，如用户缺少关注列表等。去除不完整记录后共有 1 039 名用户，其中共有 332 名用户编辑了 1 871 个标签。

②中文分词

标签编辑的随意性使得标签的规范性存在一定问题，为了更加确认单词的意思以加强它对兴趣的表征意义，需要对某些用户标签进行中文分词。本文利用 R 语言基于 ICTCLAS 中文分词系统对经过步骤①处理的标签进行分词。

该系统在中文分词中准确度较高，具有新词识别、添加新词等功能。能够自动识别新词，用户也可以根据需要添加新词，以提高分词的准确性，例如对"科幻电影"、"爱情电影"等继续分词将干扰后续计算的词定义为新词，使其不再进一步拆分，提高了样本分词准确性。经过分词总共可以得到 1 500 个分词，词频总数为 3 510，部分结果如表 2 所示。

表 2　标签分词词频统计

标签	旅游	美食	时尚	生活	新闻	后	电影	音乐	笑	…
词频	57	48	40	38	34	31	31	29	28	…
权重 w/%	1.6239	1.3675	1.1396	1.0826	0.9687	0.8832	0.8832	0.8262	0.7977	…

③去停用词

经过分词后的标签中有一部分是没有意义的，如阿、座、一定、后、有、笑等。这些停用词对研究的关系不大，通过停用词表予以去除。利用 R 语言进行停用词去除，共得到 1 281 个分词，词频总数为 2 801，部分结果如表 3 所示。

表 3　标签去停用词词频统计

标签	旅游	美食	时尚	生活	新闻	电影	音乐	娱乐	搞笑	…
词频	57	48	40	38	34	31	29	27	26	…
权重 w/%	2.035	1.7137	1.4281	1.3567	1.2139	1.1067	1.0353	0.9639	0.9282	…

④语义映射

经过以上处理后的部分标签还存在标签语义问题，如旅游和旅行、信息与资讯等，本文根据《同义词词林》，利用 R 语言计算标签间的语义相似度，以达到标签规范化的目的，提升其后分析的准确性，部分结果如表 4 所示。

表 4　标签语义映射词频统计

标签	旅游	美食	搞笑	音乐	时尚	生活	新闻	电影	娱乐	…
词频	80	48	48	42	40	38	34	31	27	…
权重 w/%	2.8633	1.718	1.718	1.5032	1.4316	1.3601	1.2527	1.1095	0.9664	…

3　基于用户标签及关注的推荐模型

3.1　用户标签模型

首先根据用户标签信息，将用户标签转换成向量并形成用户标签矩阵，根据两个用户的标签分词后相同的词语越多，则两个用户样本距离越近的原理，通过距离计算得到基于标签的用户间的距离，为后续研究做准备。

(1) 向量表示

选取预处理后标签词频大于 2 的标签(共 387 个)作为标签集 L，对用户分词后的标签进行向量化表示。数据集 D 中共 332 名用户分别将分词后的标签与 L 中的标签进行匹配，若存在即记为 1，不存在则为 0，构建矩阵，部分数据如表 5 所示。第一列为用户，每名用户以“U+ID”的形式加以区分；第一行为用户标签。

表 5　用户标签矩阵

用户	旅游	美食	搞笑	音乐	时尚	生活	新闻	电影	娱乐	…
U5107361689	1	0	0	0	0	0	1	0	0	…
U1662055430	0	0	0	1	0	0	0	1	1	…
…	…	…	…	…	…	…	…	…	…	…
U1654603903	1	1	0	0	1	0	0	1	1	…
U1692712653	1	0	0	1	0	0	0	0	1	…
U1651891204	1	0	0	0	1	0	0	0	0	…
…	…	…	…	…	…	…	…	…	…	…
U3524931687	0	1	0	0	0	1	1	0	0	…
U2040810221	1	1	0	0	1	0	1	0	0	…
U1215144691	1	1	0	1	1	0	0	1	0	…
U2684123023	0	1	0	1	1	0	0	1	0	…
…	…	…	…	…	…	…	…	…	…	…

(2) 用户间距离矩阵

对表 5 中的矩阵做用户间距离的计算[14]，设用户向量为:

$$x_i = (\delta_i(1,l), \delta_i(2,l), \cdots, \delta_i(m,l))^T, i = 1,2,\cdots N \quad (1)$$

其中，N 为样本用户数量，m 为标签集 L 中标签，l 表示第 m 个标签下的值。

$$\delta_i(m,l) = \begin{cases} 1, & 表示用户i存在标签m \\ 0, & 表示用户i不存在标签m \end{cases} \quad (2)$$

设有两个用户 x_i 和 x_j，若 $\delta_i(m,l) = \delta_j(m,l) = 1$，则称这两个用户在第 m 个标签上 1-1 配对；若 $\delta_i(m,l) = \delta_j(m,l) = 0$，则称这两个用户在第 m 个标签上 0-0 配对；若 $\delta_i(m,l) \neq \delta_j(m,l)$，则称这两个用户在第 m 个标签上不配对。记 n_1 为 x_i 和 x_j 在 m 个标签中 1-1 配对总数，n_0 为 x_i 和 x_j 在 m 个标签中 0-0 配对总数，n_2 为不配对总数，则有：$n_0+n_1+n_2=m$，用户 x_i 和 x_j 之间的距离定义为:

$$d_{ij} = \frac{n_2}{n_1 + n_2} \quad (3)$$

根据公式(3)利用 R 语言求得所有用户间的距离，部分数据如表 6 所示。

通过表 6 可以看出不同用户间的距离有所不同，d_{ij} 值越大说明两用户间距离越大，两者标签相似度越低；相反，d_{ij} 值越小说明两用户间距离越小，两者标签相似程度越高。但标签仅仅能代表用户相对静态的特征，不能及时表征用户的动态兴趣，因此本文提出在此基础上构建用户关注模型。

3.2　用户关注模型

首先根据用户关注信息，选取少量用户探究用户之间的关注关系，进而将用户关注转换成向量并形成用户关注矩阵，根据两个用户相同的关注用户越多，则两个用户样本距离越近的原理，通过距离计算得到基于标签的用户间的距离，为后续研究做准备。

表 6　基于标签的用户间距离矩阵

d_{ij}	U1	…	U80	…	U160	U161	…	U240	…	U332
U1	0	…	0.875	…	0.777778	0.818182	…	0.9375	…	0.909091
U2	0.9	…	0.9	…	0.666667	0.5	…	0.75	…	0.75
…	…	…	…	…	…	…	…	…	…	…
U80	0.875	…	0	…	0.777778	0.818182	…	0.9375	…	0.8
…	…	…	…	…	…	…	…	…	…	…
U160	0.777778	…	0.777778	…	0	0.5	…	0.888889	…	0.75
U161	0.818182	…	0.818182	…	0.5	0	…	0.8	…	0.833333
…	…	…	…	…	…	…	…	…	…	…
U240	0.9375	…	0.9375	…	0.888889	0.8	…	0	…	0.777778
…	…	…	…	…	…	…	…	…	…	…
U332	0.909091	…	0.8	…	0.75	0.833333	…	0.777778	…	0

(1) 用户共同关注关系挖掘

为了探究用户之间的关注是否存在关系，从全部332 名用户数据中随机选取 15 名样本用户的关注列表, 15 名用户关注数据如表 7 所示。

表 7　15 名用户关注列表

用户	关注列表
U3694919990	5186027114, 5182575519…
U3948635268	1642630543, 5982981128…
U3323442082	5186027114, 3440325930…
U2155768741	3766659924, 3752852352…
U3524931687	2997829562, 5611200000…
U1990226474	5878659096, 5768117490…
U1108476625	5991719510, 2781627392…
U3175953062	2705706381, 3003417253…
U2912473701	5357651574, 2415848337…
U1288915263	3937348351, 1289945134…
U2029728883	5785953533, 3174322363…
U5177961014	5796731205, 1999607273…
U2206498342	2703907413, 5465835912…
U3101945993	5980283108, 5980023345…
U5721022666	5581785513, 2850809427…

每名用户只要关注一个其他用户，则与该用户构成关注关系，15 名用户共关注 2 188 名用户，即得到2 188 个关注关系。通过 Gephi 软件对用户间的关系进行挖掘[15]，以证明基于用户关注关系聚类的可行性，如图 2 所示。

图 2 中每个用户群的中心点代表不同的中心用户，发散的点代表其关注的用户，可以看出许多中心用户关注的用户有较大的重合部分，即不同用户群之间的连线表明两个中心用户间有共同关注的对象，正是由于不同用户间存在共同关注的对象，因此用户节点数为 1 929，即 15 名用户共同关注了 259 名其他用户，同时颜色越相近的用户群则中心用户间共同关注的用户越多。根据对 15 名样本用户关系的验证，可以得出全部用户之间存在非常密切的关注联系，这对全部用户的关注关系进行聚类有重要的意义。

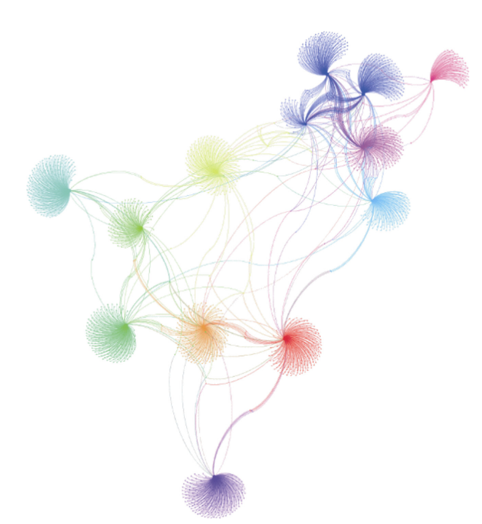


图 2　用户关注图

(2) 向量表示

将数据集 D 中共 332 名用户的关注列表进行整理，共有关注 26 958 个，删除重复关注 5 155 个，剩余关注21 803 个。将 21 803 个关注 ID 作为关注集 F，分别将每名用户的关注列表与 F 中的关注进行匹配，若存在即记为 1，不存在则为 0，构建矩阵。共 332 行用户行，21 803 列关注列，部分数据如表 8 所示。

表 8 用户关注矩阵

用户	F5186027114	…	F5608272697	F3756087501	…	F2803301701	F2516014697	…
U1846588483	1	…	0	0	…	0	0	…
U2542011901	1	…	0	0	…	0	0	…
…	…	…	…	…	…	…	…	…
U1692712653	1	…	0	0	…	0	1	…
U1644572034	1	…	0	0	…	0	0	…
U1781457455	0	…	0	0	…	0	0	…
…	…	…	…	…	…	…	…	…
U5107361689	0	…	0	0	…	0	0	…
U2542011901	1	…	0	0	…	0	0	…
U2834863492	0	…	1	1	…	1	1	…
…	…	…	…	…	…	…	…	…
U3524931687	0	…	1	1	…	1	0	…
U1203156407	0	…	0	1	…	0	0	…
…	…	…	…	…	…	…	…	…

(3) 用户间距离矩阵

采用与标签距离计算同样的算法计算用户间距离，得到基于关注关系的用户间距离矩阵，部分数据如表 9 所示。

表 9 基于关注关系的用户间距离矩阵

d_{ij}	U1	…	U80	…	U160	U161	…	U240	…	U332
U1	0	…	0.963350	…	0.988636	…	…	0.970149	…	0.991701
U2	0.994350	…	0.992753	…	1	0.993827	…	1	…	1
…	…	…	…	…	…	…	…	…	…	…
U80	0.963350	…	0	…	0.994680	0.994186	…	0.991525	…	0.997076
…	…	…	…	…	…	…	…	…	…	…
U160	0.988636	…	0.994680	…	0	0.995762	…	0.992187	…	0.987012
U161	0.987654	…	0.994186	…	0.995762	0	…	0.996491	…	0.989664
…	…	…	…	…	…	…	…	…	…	…
U240	0.970149	…	0.991525	…	0.992187	0.996491	…	0	…	0.992882
…	…	…	…	…	…	…	…	…	…	…
U332	0.991701	…	0.997076	…	0.987012	0.989664	…	0.992882	…	0

根据表 9，d_{ij} 越大说明两用户间关注的相似度越低，d_{ij} 越小说明两用户间关注的相似度越高。同时可以看到距离矩阵中有一部分值是 1，这是因为关注集 F 中 21 803 个关注 ID 相对于用户最多 200 的关注过于庞大，造成数据的稀疏性。由此可以发现若仅根据用户关注对用户进行聚类实现个性化推荐还是有一定的缺陷的。

3.3 综合用户聚类模型

将用户标签静态性与用户关注的动态性进行综合聚类。利用多维尺度分析法对多维度的用户标签与用户关注进行降维后，再通过 K-means 方法进行用户聚类，实现用户的个性化推荐。

(1) 向量表示

多维尺度分析法(MDS)[16-17]是一种将多变量的多维大型数据压缩到低维空间的方法，通过低维空间的点表示变量间的潜在规律性联系，且通过平面间的距离反映样本间的相似度。MDS 具有很多优点，包括[10]：样本数据可以不受任何事先分布假设的约束；能够处理不同类型的数据；能够将多变量多维数据压缩到低

维空间等。

本文根据用户标签及关注关系的向量矩阵，分别对其进行 MDS 降维处理，将维数差别巨大的标签矩阵(332 行×387 列)与关注矩阵(332 行×21807 列)信息整合到二维空间中，用户的标签 MDS 和关注 MDS 代表用户在向量空间中的维度，其值为用户在向量空间中的坐标，部分结果如表 10 所示。

表 10　基于标签及用户关注 MDS 降维数据

用户	标签 MDS	关注 MDS
U2612101423	0.049094493	–0.034319904
U1846588483	0.014763293	–0.011171253
U1306794125	0.055376563	–0.034743694
U5179732445	0.50130544	–0.036149048
U5761248787	0.50130544	–0.004671656
U1665102492	0.04820318	–0.033469629
U2647197351	0.033225349	–0.046390183
U5961019705	0.034749234	–0.03427661
U1781457455	0.043747374	–0.034271488
U5107361689	–0.055230674	0.114665726
U2542011901	0.046136223	–0.000205833
U2871542364	0.058303826	–0.042518174
U2834863492	0.05151389	0.004734437
U2624882007	–0.081583674	–0.027694683
U1692712653	–0.08441402	–0.004928777
U1644572034	0.052114494	0.095748648
U1651891204	–0.139576002	–0.029852541
U2094215167	0.050809285	0.003524086
U3524931687	–0.10443334	–0.023421971
…	…	…

(2) 用户聚类

K-means 算法是一种典型的适合于大样本的 Q 型聚类分析方法[14]，通过计算数据集中点与点之间的距离或相似度进行聚类，且类中心采用类中值的均值计算而成[18]。

聚类算法如下：

①确定初始类中心点。随机选择 k 个元素作为 k 个类的中心点。

②初始类。将表 5 和表 8 中基于标签和用户关注 MDS 降维数据结合，计算每个点到类中心的距离，将每个点聚类到离该点最近的类中去，得到 k 个粗分类。

③更新类中心。计算各个粗分类中所有点的坐标平均值，并将这个平均值作为新的聚类中心。

④重复执行步骤②、步骤③，直到聚类中心不再进行大范围移动。

K-means 聚类作为凝聚式的聚类方法，需要人为定义其初始类中心点的个数，由于样本数据共有 332 名有效用户，为不失一般性，模型为每位用户推荐 10 名左右的用户，因此以初始类中心 k=30 为例进行聚类，聚类结果如图 3 所示。

图 3 中不同形状的点表示不同的用户簇群。米字型代表簇中心所在的位置，簇中心为该簇中所有用户坐标的平均值，该中心点即代表该簇，用以表征该簇中的所有用户。可以看到，每一个簇中心周围都聚集着该簇中的点，且较为紧密，与其他簇中心有较为明显的距离，这说明聚类效果较好。但仍需通过聚类指标对聚类效果进行衡量，表 11 为综合聚类结果的指标。

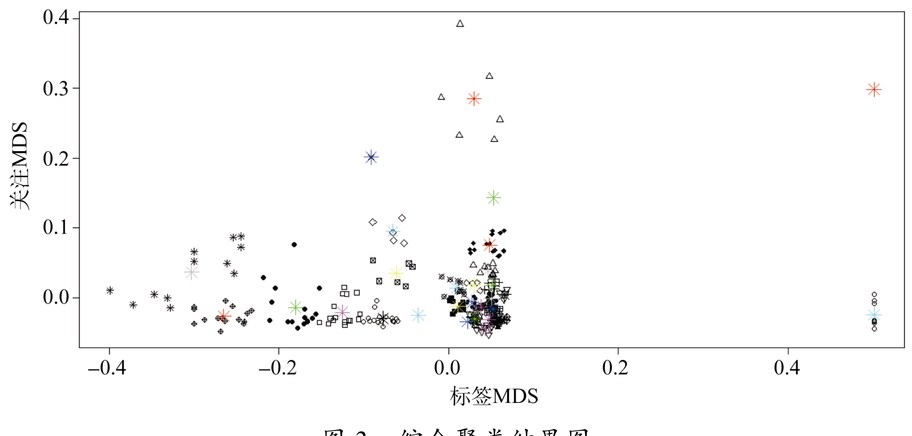


图 3　综合聚类结果图

表 11　综合聚类 k=30 指标结果

指标	值
TOT.Withinss 簇群内距离平方总和	0.1385733
Betweenss 簇群间距离平方总和	7.615879

如表 11 所示，簇群内距离平方总和(TOT.Withinss)指标表示所有簇用户距离其簇中心点距离平方的和，该指标用以衡量聚类结果的凝聚度，该值越小说明该类用户越紧凑，聚类效果越好；簇群间聚类平方总和(Betweenss)指标表示不同簇群间簇中心距离的平方和，该指标用以衡量聚类的分离度，该值越大说明将类与类之间分离越明显，聚类效果越好。

4　模型效果分析

4.1　模型有效性评价

(1) 评价指标

由于聚类分析是一种无监督的分析方法[18]，因此对聚类后的结构进行有效性度量是非常必要的。聚类有效性的度量一般基于对簇内和簇间两个方面进行衡量，好的聚类效果为具有最小的簇内距离和最大的簇间距离，即具有最小的簇内凝聚度和最大的簇间分离度[7]。

当前提出的有效性函数大多是基于凝聚度和分离度的组合进行改进。Xie-Beni 提出使用 V_{XB} 函数对聚类有效性进行测量[19-20]，如公式(4)所示。

$$V_{XB}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|v_i-x_j\|^2}{n\cdot\min\|v_i-v_j\|^2} \tag{4}$$

其中，V_{XB} 表示凝聚度和分离度的比例，V_{XB} 越小说明聚类效果越好；$\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|v_i-x_j\|^2$ 为度量凝聚度，其值越小该类越紧凑；$\min\|v_i-v_j\|^2$ 为度量分离度，其值越大，分离度越大，则类与类之间分离得越好。

本文将上述函数简化，如公式(5)所示。

$$V_{TB}=\frac{TOT.Withinss(k)}{Betweenss(k)} \tag{5}$$

其中，k 表示聚类数，*Tot.Withnss*(k)表示在聚类数为 k 下，簇内距离平方和总量，用以度量凝聚度；*Betweenss* 表示在聚类数 k 下，簇间聚类平方和总量，用以度量分离度，V_{TB} 值越小，则聚类效果越好。

(2) 有效性分析

为了方便描述，将本文提出的基于标签与关注关系综合聚类方法简写为 L_F_C；将基于标签的聚类方法简写成 L_C；将基于关注聚类的方法简写成 F_C。使用本文提出的 V_{XB} 函数的简化函数 V_{TB} 函数。分别预设聚类个数，这里设定各方法聚类个数均为 k=3、k=6、k=9、k=12、k=15、k=18、k=21、k=24、k=27、k=30，根据标签距离矩阵、关注距离矩阵及综合 MDS 矩阵分别经过聚类并计算得到图 4。

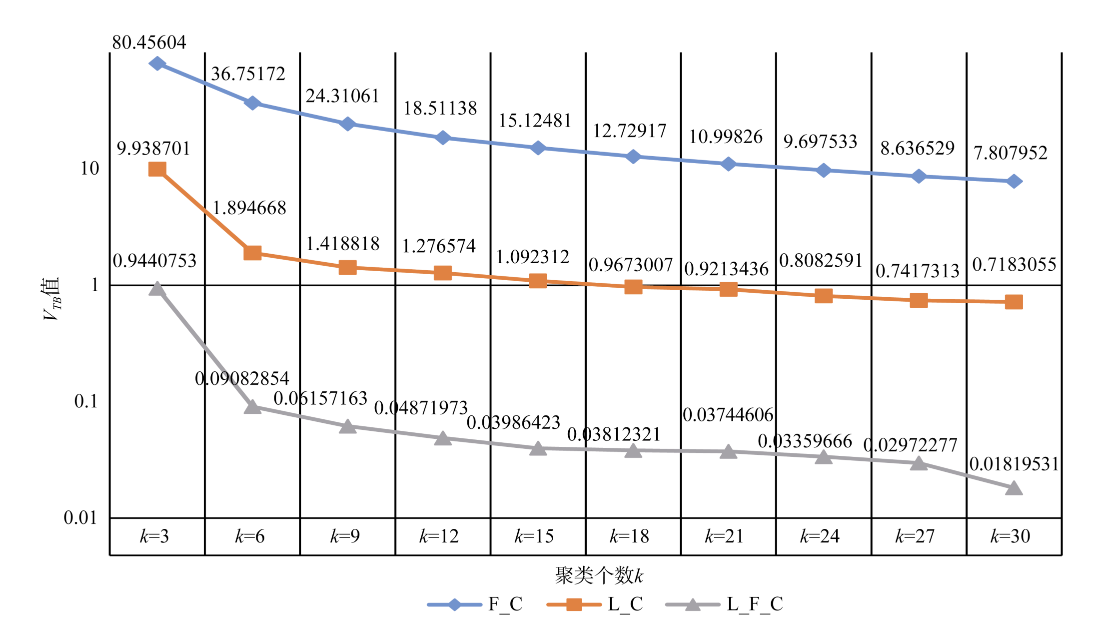


图 4　L_F_C、F_C 和 L_C 方法 V_{TB} 值对比

从图 4 可以看出本文提出的基于标签及关注关系综合聚类(L_F_C)在 V_{TB} 指标上远远优于单独基于标签聚类方法(L_C)和基于关注关系聚类方法(F_C)。表明本文所提出的基于标签及关注关系聚类的方法能够获取较好的聚类结果。主要原因在于 L_F_C 方法将用户静态标签及用户动态关注关系考虑其中，大大增加了聚类的准确性及有效性。

4.2 实证结果分析

本文随机选取用户 M 对模型进行实证。用户 M 数据如表 12 所示。

设置聚类数 k=30 对样本 332 名用户数据进行聚类，聚类结果如表 13 所示。

表 12 用户 M 数据

用户 ID	用户昵称	标签	关注列表
2132089917	陈秋实和他的朋友们	语录，新闻，美剧，运动，80 后，传媒，写作，处女座	1803526210, 1854768217, …

表 13 模型聚类结果

用户 ID	用户昵称	标签	关注列表
2132089917	陈秋实和他的朋友们	语录，新闻，美剧，运动，80 后，传媒，写作，处女座	1803526210, 1854768217, …
1448466905	非要马甲线	下厨房，营养学，健身，爱，天蝎，美食，旅游	1690832323, 1238296465, …
1592611830	演员李健	天蝎座	1870958692, 5941080382, …
2307134004	STAGExx	时尚，美食，音乐，电影，旅游	1813787671, 1812640242, …
3173913704	葡萄 sasa 定制店	旅游，时尚	5646244946, 3944457562, …
1254995044	山外有	电脑，宅，书，纪录片，摄影，西南交通大学，四川大学	64230524, 3208535250, …

通过对用户 M(陈秋实和他的朋友们)背景进行了解，可以发现该用户昵称叫陈秋实，是《我是演说家》亚军，从事过演员助理、配音员、记者、电视编导、电视主持人、舞台剧、影视剧演员等多种职业，目前就职于北京隆安律师事务所，主要执业方向为影视娱乐、传媒、互联网领域的法律业务。

因此，用户 M 对影视、传媒、互联网等行业应较为关注，从表 12 可以发现虽然该用户在标签中并未明确标注“娱乐”、“互联网”等词语，但对用户 M 的推荐主要是娱乐、互联网领域的用户，可以从“演员李健”、“STAGExx”等用户的标签中发现。同时从演员李健的标签中也可以看出，演员李健标签只有“天蝎座”，但其身份为一名演员，模型通过关注关系发现该用户的潜在特征，将其推荐给用户 M。

同时，经过对用户 M 关注列表的分析，该用户在最近关注了“享骑出行”等出行旅游类微博，因此模型也将基于关注关系为用户 M 进行推荐。根据推荐结果可以发现，虽然用户 M 在标签中并未有“旅游”等词语，但在其推荐用户中可以看到“非要马甲线”、“STAGExx”、“葡萄 sasa 定制店”三名用户的标签中都含有“旅游”标签，说明这三者都是对旅游出行具有长期兴趣的用户，模型对用户 M 关注关系的更新发现他们与用户 M 关系，进而进行推荐。

综上所述，本文所提出的模型综合用户 M 标签表征的长期兴趣与关注表征的短期兴趣能够较好地将符合用户 M 特征的其他用户作为被推荐对象，推荐给用户 M。但是，由于样本信息不完全，主要集中在娱乐领域，因此，在被推荐用户中法律领域的用户并未出现。经过上述分析有理由相信，在数据量更为充分的情况下，模型将能更精确地综合用户长短期兴趣，推荐更为准确的相似用户。

5 结 语

本文将用户作为个性化推荐的对象，提出基于用户静态标签与动态关系网络的用户推荐模型。通过用户标签及用户关系网络挖掘用户长短期兴趣特征，开创性地利用 MDS 降维的方式将用户多维信息全部包含进模型中，并使用聚类分析的方法发现潜在相似用户，提高了用户聚类的准确性与全面性及用户推荐的有效性。并且，本文将提出的模型应用于真实数据集，证明了模型的准确性及推荐的有效性。

但本文为了更加清晰地描述模型，并未从多个角

度进行数据的采集，样本数据集具有局限性，不能完全涵盖用户所有兴趣领域，仅从一个领域验证了模型的准确性与有效性。今后的研究方向将扩大数据的覆盖面，从多个领域节点出发收集数据，通过实证结果继续完善模型的相关算法，以进一步提高模型的可行性和有效性，促使模型从理论走向实践。

参考文献：

[1] 熊回香，王学东. 大众分类体系中标签概念空间的构建研究[J]. 情报学报，2012，31(9)：984-992. (Xiong Huixiang, Wang Xuedong. Research on Tag Concept Space Construction in Folksonom System[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(9): 984-992.)

[2] 熊回香，杨雪萍. 社会化标注系统中的个性化信息推荐研究[J]. 情报学报，2016，35(5)：549-560. (Xiong Huixiang, Yang Xueping. Personalized Information Recommendation Research Based on Combined Condition in Folksonomies[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(5): 549-560.)

[3] Arekar T, Sonar M R S, Uke N J. A Survey on Recommendation System[J]. IOSR Journal of Computer Engineering, 2015, 5(1): 1-4.

[4] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(6): 734-749.

[5] 何晓林. 基于用户兴趣学习的个性化信息服务模型研究[D]. 北京：北京交通大学，2008. (He Xiaolin. Research on Personalized Information Service Model Based on User Interest Study[D]. Beijing: Beijing Jiaotong University, 2008.)

[6] 易明，操玉杰，沈劲枝，等. 社会化标签系统中基于密度聚类的 Web 用户兴趣建模方法[J]. 情报学报，2011，30(1)：37-43. (Yi Ming, Cao Yujie, Shen Jinzhi, et al. An Approach to Web User Interest Modeling Based on Density-based Clustering Algorithm in the Social Tag System[J]. Journal of the China Society for Scientific and Technical Information, 2012, 30(1): 37-43.)

[7] 王向前，李慧宗. 基于资源内容聚类的社会化标签聚类方法[J]. 情报杂志，2016，35(11)：141-145. (Wang Xiangqian, Li Huizong. A Method of Tag Clustering Based on Resource Contents[J]. Journal of Intelligence, 2016, 35(11): 141-145.)

[8] Shepitsen A, Gemmell J, Mobasher B, et al. Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering[C]//Proceedings of the 2008 ACM Conference on Recommender Systems, 2008: 259-266.

[9] Gemmell J, Shepitsen A, Mobasher B, et al. Personalizing Navigation in Folksonomies Using Hierarchical Tag Clustering[C]//Proceedings of International Conference on Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, 2008: 196-205.

[10] 卢小宾，孟玺，张进. 基于词共现的社会化标签研究热点可视化分析[J]. 情报学报，2012，31(2)：204-212. (Lu Xiaobin, Meng Xi, Zhang Jin. Visualization of Hot Topics in Social Tagging Based on Co-words Analysis Method[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(2): 204-212.)

[11] 柴彦. 基于共词聚类分析方法的知识管理国内研究述评[J]. 情报科学，2015，33(4)：149-153. (Chai Yan. Review of Knowledge Management Based on Co-Word Clustering Analysis[J]. Information Science, 2015, 33(4): 149-153.)

[12] Masnick A M, Valenti S S, Cox B D, et al. A Multidimensional Scaling Analysis of Students' Attitudes about Science Careers[J]. International Journal of Science Education, 2010, 32(5): 653-667.

[13] 黄红霞，章成志. 中文微博用户标签的调查分析——以新浪微博为例[J]. 现代图书情报技术，2012(10)：49-54. (Huang Hongxia, Zhang Chengzhi. Investigation and Analysis of Chinese Microblog UserTags——Using Sina Weibo as Example [J]. New Technology of Library and Information Service, 2012(10): 49-54.)

[14] 薛毅，陈立萍. 统计建模与 R 软件[M]. 第 1 版. 北京：清华大学出版社，2007. (Xue Yi, Chen Liping. Statistical Modeling and R Software[M]. The 1st Edition. Beijing: Tsinghua University Press, 2007.)

[15] Cherven K. Network Graph Analysis and Visualization with Gephi[M]. Packt Publishing, 2013.

[16] 郭婷，郑颖. 数据挖掘在国内图书情报领域的应用现状分析——基于文献计量分析和共词分析[J]. 情报科学，2015，33(10)：91-98. (Guo Ting, Zheng Ying. Research on the Application of Data Mining in the Field of Library and Information Science in China——Based on Bibliometric Analysis and Co-word Analysis[J]. Information Science, 2015, 33(10): 91-98.)

[17] Wikipedia. Multidimensional Scaling[EB/OL]. [2016-11-01]. https://en.wikipedia.org/wiki/Multidimensional_scaling.

[18] Harrington P. 机器学习实战[M]. 曲亚东，李锐，王斌等译. 第 1 版. 北京：人民邮电出版社，2013：184-185. (Harrington P. Machine Learning in Action[M]. Translated by Qu Yadong, Li Rui, Wang Bin, et al. The 1st Edition. Beijing:

Posts & Telecom Press, 2013: 184-185.)

[19] 张宇献，刘通，董晓，等. 基于改进划分系数的模糊聚类有效性函数[J]. 沈阳工业大学学报，2014，36(4)：431-435. (Zhang Yuxian, Liu Tong, Dong Xiao, et al. Validity Function for Fuzzy Clustering Based on Improved Partition Coefficient[J]. Journal of Shenyang University of Technology, 2014, 36(4): 431-435.)

[20] 朱连江，马炳先，赵学泉. 基于轮廓系数的聚类有效性分析[J]. 计算机应用，2010，30(S2)：139-141. (Zhu Lianjiang, Ma Bingxian, Zhao Xuequan. Clustering Validity Analysis Based on Silhouette Coefficient[J]. Journal of Computer Applications, 2010, 30(S2): 139-141.)

作者贡献声明:

熊回香：提出研究方向和方法，论文撰写指导，论文修订；
蒋武轩：数据获取，数据分析，论文撰写，论文修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 412370630@qq.com。

[1] 熊回香，蒋武轩. 实验数据及数据预处理数据.xlsx. 微博实验数据.

[2] 熊回香，蒋武轩. 实验数据及数据预处理数据.xlsx. 分词词频结果数据.

[3] 熊回香，蒋武轩. 实验数据及数据预处理数据.xlsx. 去停用词词频结果数据.

[4] 熊回香，蒋武轩. 实验数据及数据预处理数据.xlsx. 标签语义映射词频数据.

[5] 熊回香，蒋武轩. 用户标签处理过程数据.xlsx. 用户标签矩阵.

[6] 熊回香，蒋武轩. 用户标签处理过程数据.xlsx. 基于用户标签的用户间距离矩阵.

[7] 熊回香，蒋武轩. 用户关注关系处理过程数据.xlsx. 15 名样本用户关注列表.

[8] 熊回香，蒋武轩. 用户关注关系处理过程数据.xlsx. 用户关注矩阵.

[9] 熊回香，蒋武轩. 用户关注关系处理过程数据.xlsx. 基于关注关系的用户间距离矩阵.

[10] 熊回香，蒋武轩. 综合用户聚类过程数据及验证用户 M 数据.xlsx. 基于标签及用户关注 MDS 降维数据.

[11] 熊回香，蒋武轩. 综合用户聚类过程数据及验证用户 M 数据.xlsx. 综合聚类结果.

[12] 熊回香，蒋武轩. 综合用户聚类过程数据及验证用户 M 数据.xlsx. 模型评价.

[13] 熊回香，蒋武轩. 综合用户聚类过程数据及验证用户 M 数据.xlsx. 用户 M 推荐结果.

收稿日期: 2017-04-07
收修改稿日期: 2017-05-15

Clustering and Recommending Users Based on Tags and Relation Network

Xiong Huixiang Jiang Wuxuan
(School of Information Management, Central China Normal University, Wuhan 430079, China)

Abstract: [**Objective**] This paper proposes a new model to recommend potential similar users with the help of social tags and relation network. [**Methods**] First, we explored characteristics of the users' short or long-term interests based on the social tagging system. Then, we built a user-clustering model using multidimensional scaling method with the tags and relationship data. Finally, we recommended similar users based on the clustering results. The proposed model was examined with Weibo data. [**Results**] We found that the new model could effectively combine the characteristics of the user's interests, and then identify the potential similar ones. [**Limitations**] The sample data does not include everything on user interests. Thus, we only examined the effectiveness of the proposed model with limited data. [**Conclusions**] The user recommendation model based on static tags and dynamic relational network could improve the personalized recommendation services.

Keywords: Social Tagging Tag Relation Network User-cluster Multidimensional Scaling Analysis